文章编号:1000-2367(2022)05-0063-08 DOI:10.16366/j.cnki.1000-2367.2022.05.009

知乎标签网络的演化及模型研究

黄涛[1a],王胜烽[1b],吴晔[2],张鹏[1a],肖井华[1a]

(1.北京邮电大学 a.理学院;b.信息与通信工程学院,北京 100876;
2.北京师范大学 计算传播学研究中心,广东 珠海 519087)

摘 要:知识网络是探索知识发展脉络和形成机制的重要基础.研究知识网络的统计特性和形成机理具有重要意义.标签网络作为知识网络的一种,近年来受到研究者的关注,但是目前关于标签网络生长机制的研究还较为缺乏.基于知乎问答平台的标签数据构建知识标签网络,统计分析了标签网络的静态统计特性以及演化特性.为了理解知识标签网络复杂结构的形成原因,提出了知识激发问题的网络动态生长模型,模型假定新问题由知识标签激发,知识标签激发问题的能力与其度值正相关.仿真结果表明,模型可以很好地再现知乎标签网络的统计特性和社团化结构.研究结果揭示了标签知识网络增长过程中表现的动态演化特性,并基于实证结果建立的标签网络生长模型,对理解知识网络的形成和发展有一定启发意义.

关键词:标签网络;问答社区;演化模型;幂律分布

中图分类号:O157.5 **文献标志码**:A

知识是人类认知世界的成果,是社会化合作进程的产物[1].知识网络是知识体系的具象化,其结构特征和演化机制是探究知识发展脉络和创新趋势的基础,对研究知识的发展与创新具有重要意义[2].知识网络从研究内容上可以分为两类,一是以科学文献、专业知识为主要研究对象的专家知识网络,比如引证网络[3-4]、关键词网络[5-6]、合著网络[7]等;另一类是以互联网信息、大众知识为主要研究对象的大众知识网络,如问答平台标签网络[8]、Wiki 知识网络[9]、博客交流网络[10]等.

已有的研究证实,专家知识网络具有复杂的网络结构.如徐汉青等[11]利用学术网站 CiteUlike 上的数据构建领域知识网络,分析得出网络始终保持小世界特性并逐步趋近于稳定的无标度网络.耿志杰等[12]获取 CSSCI 数据库中情报学领域的期刊文献数据,基于关键词的共现关系构建关键词网络,运用复杂网络分析方法发现关键词知识网络是典型的无标度网络,并且满足小世界特性.滕广青等[13]基于中国知网的文献数据和文献标注系统中的文献数据分别构建关键词知识网络和标签知识网络,并通过对原始网络和提取的层次知识网络分析,发现无论是关键词知识网络还是标签知识网络,都具备无标度特性和小世界特性.

大众知识网络涉及的知识主题广泛全面,其网络构建取决于拥有不同教育背景,思想观念和行为的全部参与者,所以大众知识网络也拥有复杂的网络结构.有很多学者就大众知识网络的网络结构展开了分析研究,如潘旭伟等[14]以 Wikipedia 为研究对象,结合复杂网络分析方法,对构建的 Wiki 词条和主题参考网络实证分析,结果表明 Wiki 知识网络的入度服从幂律分布,网络具有小世界效应.标签知识网络作为大众知识网络的一种,其本质是一种共现网络[15].网络基础数据是 Folksonomy 模式下生成的文字标签.这些标签数据被认为包含了大量用户的思维观点、行为特征和喜爱偏好,而且这些数据相结合能够生成新颖的知识和见解[16].对标签知识网络的研究能够帮助把握大众知识体系的发展趋势,推动知识的创新.已有的很多文献(如[17-18])揭示了在标签知识网络中也存在幂律分布主导网络的情况,也就是网络具有无标度特性.

收稿日期:2021-12-18;**修回日期**:2022-03-01.

基金项目:国家重点研发计划(2020YFF0305300)

作者简介:黄涛(1995-),男,湖南衡阳人,北京邮电大学硕士研究生,研究方向为网络科学,E-mail:huangtao125@bupt.edu.cn.

通信作者:王胜烽,E-mail:docshengfeng@foxmail.com;肖井华,E-mail:jhxiao@bupt.edu.cn.

除了静态结构分析以外,作为一种知识网络,标签知识网络的演化特征和演化机制的研究也逐渐得到了关注.有学者基于复杂模体分析了标签网络的演化特征,得出问答平台上的知识标签网络在网络结构上由趋于稳定的态势[19].BARABÁSI 等人[20]提出的 BA 模型被用于解释知识网络的无标度特性.模型中新节点优先连接度值大的节点,择优概率一般采用 $\Pi(k_i)=g(k_i)/\sum g(k_i)$,$k_i$ 表示节点的度值,经典 BA 模型中 $g(k_i)=k_i$,为线性择优[20].后来学者们研究发现在一些实际网络中并不一定是线性择优,所以对经典 BA 模型进行了改进,如提出 $g(k_i)=k_i^r$ 的非线性择优模型[21],当 $r>1$ 时为超线性相关[22],当 $1>r>0$ 时为亚线性相关[23].关于标签知识网络的演化机制研究方面,韩仪等人在经典 BA 模型上引入“批量增长”和“交叉连接”特性[24],对知识标签网络社团形成进行了解释.

综合上述研究,知识网络具有无标度和社团共存的宏观特性,BA 网络模型被用于解释其形成原因.尽管类似 BA 的优先增长规律被广泛认可,但是对于标签网络的增长在哪个层面遵循此规律还需要进一步探索.在标签知识网络的形成中,BA 网络模型中每一个新节点的加入作为网络的增长周期.这并不符合所有的实际标签知识网络.例如在问答平台标签网络中,网络每一次增长都是因为新问题的出现,而且在这一过程中,新节点是否生成以及生成数目都是不确定的,所以,对于各种实际存在的标签知识网络,还需要构建从问题出发的标签模型解释标签知识网络的演化机制.

鉴于此,本文研究了中国知名在线问答平台——知乎网站上标签网络的形成机制.研究了知乎标签网络动态演化特性,如新标签的产生情况、标签之间的连接倾向情况.基于研究结果,提出了一个新的网络动态增长模型,具体上,模型以新问题的出现为网络增长周期,假定问题由知识标签激发生成,知识标签激发问题的能力与其度值正相关.模型能够很好地再现知乎知识标签网络的无标度特性和社团结构.

第一部分介绍了使用的数据集,构建并分析了标签网络的度分布和社团结构.第二部分统计分析了标签网络的动态演化特性.第三部分提出标签网络生长模型,模型生成的网络跟实际网络相符,进一步也分析了标签网络模型的参数随着时间的变化.第四部分对全文进行总结.

1 数据和标签网络

使用来自知乎网站的问题和标签数据.知乎网站是国内知名的在线问答平台,网站上的问题涉及领域非常广泛,包含了社会生活中方方面面的知识.网站上的问题都会携带至少一个标签.这些标签由用户选定,用于标记问题讨论的内容.使用的数据,其构成个体为问题.每个数据个体包括了问题的编号,问题的详细文字描述,问题创建的时间以及问题所携带的标签.经过数据清洗和筛选,研究数据包括了 656 387 个问题以及 58 632 个不同的标签.

这些问题和标签数据的时间跨度为 2011 年至 2017 年,以自然年份为时间窗口,统计每一个时间窗口内出现的问题数目和标签数目,结果如表 1 所示.知乎网站自 2013 年开放注册,之后几年发展迅速,问题和标签数目在 2015 年开始有较大增长.

表 1 时间窗口内的问题和标签的累积数目

Tab. 1 Cumulative number of issues and tags within the time windows

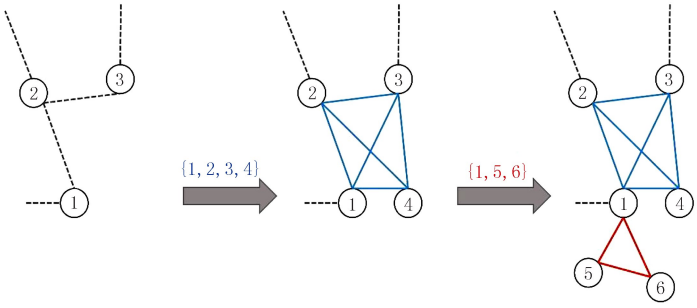
窗口年份	问题数	标签数	窗口年份	问题数	标签数
2011 年	9 957	9 701	2015 年	144 754	34 653
2012 年	16 082	11 112	2016 年	139 670	36 731
2013 年	33 167	17 024	2017 年	243 296	43 141
2014 年	69 461	25 340	总 计	656 387	58 632

标签网络将标签作为节点,以标签间的共现关系构建连边.在知乎网站中,标签的共现关系即两个标签为同一个问题携带,所以每个问题携带的标签,其两两之间都存在共现关系.如图 1,按照问题出现的时间排序,将标签及标签间的连边关系加入网络,就构建了标签知识网络.

2 标签网络的统计特性

以自然年份为时间窗口,基于每个时间窗口的问题和标签数据构建出标签知识网络,在本小节的研究

中，分别从网络度分布、社团划分对网络进行了分析，得到了知乎标签知识网络中无标度特性和社团化共存的情况.



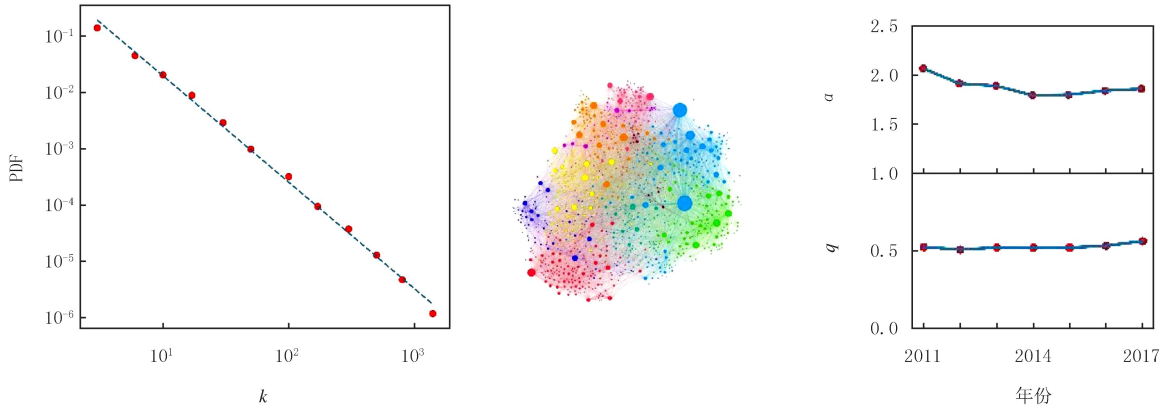
假设网站出现一个新问题({1, 2, 3, 4})，其携带标签为标签1、标签2、标签3和标签4，其中标签4为新生标签. 4个标签两两之间如无连边(如标签1和标签3)，则形成权重为1的连边；如有连边(如标签1和标签2)，则连边权重值加1. 后一时刻，网站出现新问题{1, 5, 6}，则3个标签两两之间形成权重为1的连边.

图1　知乎平台标签知识网络构建示意图

Fig. 1　Schematic diagram of Zhihu tagging network construction

从网络度分布结果上可以看出知识标签网络的加权度分布符合幂律分布.设定自然年份为时间窗口，基于时间窗口内的数据构建标签子网络.网络为有权网络，统计 2013 年标签子网络的加权度分布数据并进行拟合，结果如图 2(a)所示.拟合的幂律分布说明标签网络度的分布极不均匀，在标签网络中较少的标签和大量标签相连，说明了知乎上的问题具有很多相同的标签.统计得到各年份知乎标签网络度分布的拟合幂指数 α 较为平稳，处于 1.8～2.0 之间，见图 2(c).

现实中的很多网络都具有良好的社团结构，社团结构体现了网络的区域聚集特性.具有社团结构的网络，社团内部连接相对紧密，而社团之间的连接较为稀疏.网络社团划分所使用的指标通常是模块度.模块度的数值大小对应网络社团划分的质量，社团划分结果越好模块度越大.每年的标签子网络中节点数目众多，而大部分节点的度很小，所以截取节点度大的核心部分节点，使用 Louvain 社区划分算法，得到网络模块度最大的社团划分结果.Louvain 社区划分算法的工作过程就是动态调动节点所属的社团，以得到网络最大模块度的社团划分结果[25].图 2(b)为 2013 年子网络社区划分结果.节点大小对应节点度值，颜色相同的节点属于同一社区，点间的连边的颜色为连边两个节点颜色的混合，相同社区节点之间的紧密连接，可以形成不同颜色的社区块.可以看出知乎标签网络可以形成不同的社区块，网络模块度 q 一直能保持较高的数值，网络具有良好的社团结构.



(a)2013年子网络度分布　　(b)2013年子网络社团结构　　(c)各年子网络结构特征

图2　网络加权度分布和网络社团结构

Fig. 2　Network weighted degree distribution and network community structure

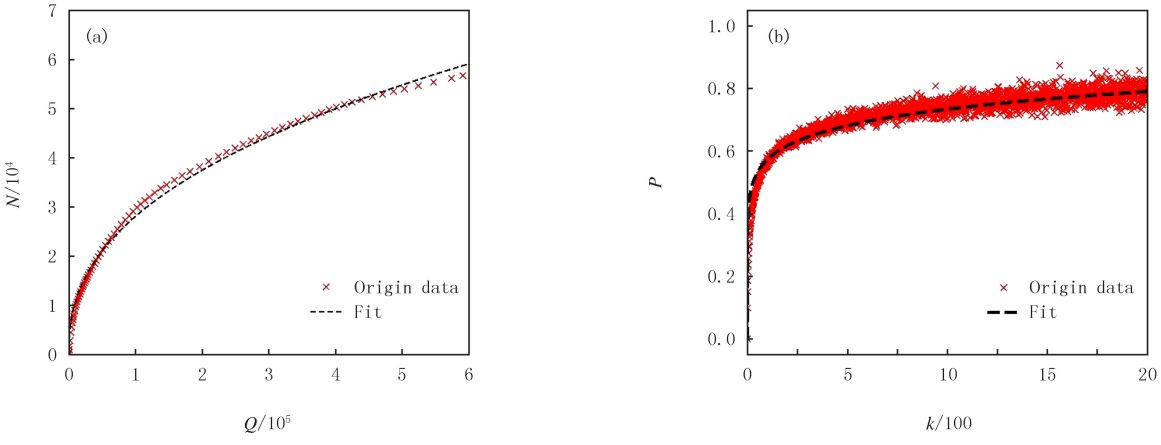
基于知乎问答平台数据构建的标签知识网络,具备无标度特性,网络还拥有良好的社团结构.网络的复杂结构特性是在其演化过程中形成的,反映了其复杂的演化机制,对于知乎标签网络的研究还需要从动态角度对其进行分析.

知乎网站的新问题是源源不断的,标签数目也会随着问题数目的增加而增加.统计所有问题携带的标签数目,汇总数目分布得到标签数 r 在 1～5 之间的占比分别为 0.246,0.159,0.184,0.167,0.244.每个问题携带标签的平均数为 $\bar{r}=3.0$.统计问题数目 Q 和相应的标签数目 N,如图 3(a),标签的增长速率在网络的演化过程中并不一样,标签数前期增长速度明显快于后期.图 3(a)中幂律分布很好地刻画标签数目和问题的关系,因此可以得到标签随问题的增长速率为

$$v=\mathrm{d}N/\mathrm{d}Q=\bar{r}(aN^{-b}),$$

其中,a 为归一化常数,(aN^{-b})表示新生问题携带的任一标签为新标签的概率取决于新生问题出现时网络已有标签总数.所以当网络中的标签数目增大时,对应的标签新生速率降低了,新问题携带新生标签的概率降低了.

当新问题的标签在已有标签网络存在时,存在连边的标签之间也会增加新的连边.与标签存在直接连边的标签为该标签的直接邻居.统计标签新增连边中连接直接邻居的占比,可以得到图 3(b).以幂函数 $P_n=ck_i^a$ 拟合,其中 c 为归一化常数.在知乎标签网络的演化中,新问题增加时,标签网络中度大的节点会进一步强化从而具有更大的度.



(a)知乎网站标签数目N随问题数目Q的增长情况.×为实际数据,来自2011年到2017年;虚线为拟合曲线$N=236.969Q^{0.415}$.
(b)统计节点的连边增量中选择连接直接邻居的占比与节点度值的对应关系以及拟合$P_n=0.350k^{0.107}$.

图3 知乎标签网络更新示意图和标签增长情况以及拟合曲线

Fig.3 Schematic diagram of label network update and label growth

3 模 型

3.1 模型构建

根据前面的分析结果,提出了一个标签网络生长模型.在线问答平台是由问题组成的,从问题生成的角度出发构建知识标签网络的演变更贴合实际情况.而每一个问题都是有核心讨论点的,问题包括问题的回答、评论等互动都是围绕核心讨论点展开的,换言之,问题的提出立足于核心讨论点,这个核心讨论点可以视为激发问题的知识标签,模型演化如图 4,模型设定标签网络变化由问题更新周期组成,每个更新周期都会产生一个新问题.新问题由父标签激发产生,相对于其他标签,新生问题更倾向于与父标签距离近的邻居标签.

在模型中,一个更新周期中,标签 i 被选为父标签的概率 P_i 与节点 i 的度值 k_i 的关系为

$$P_i=\frac{k_i^m}{\sum_j k_j^m},$$

其中,m 为度值影响因子,$m>0$.当 $0<m<1$ 时,知识标签激发问题的能力与其度值呈亚线性关系,$m=1$

时，两者呈线性关系，当然两者也可能出现超线性的关系，即 $m>1$.每个新问题具有的标签数目为 r，因此除确定了父标签外，还需要确定 $r-1$ 个其他标签，根据上一节的结果，邻居标签作为新生问题携带标签的概率与父标签的度值相关.

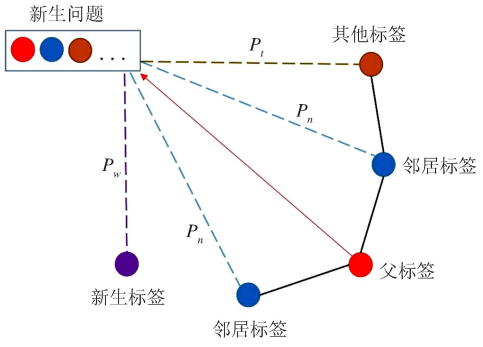
标签网络生长模型的算法如下：

1)确定新生问题的携带的标签数目 N，其中 $0<N<6$.

2)根据新生率确定新生问题携带的新生标签数目 n，其中 $0\leq n\leq 6$.新生率为已有标签数的幂函数，随已有标签数目的增大而降低.所以可能出现一个新生问题携带标签皆为新标签的情况 $P_w=aN^{-b}$.

3)以概率 P_i 选择新生问题父标签.

4)确定新生问题父标签之后，优先选择父标签的邻居标签.对每一个未确定的标签空位，会优先选择父标签的直接邻居，其选择父标签的直接邻居的概率与父标签度值 k_i 的关系 $P_n=ck_i^d$.



新生问题由父标签激发产生，其一定携带父标签，其携带邻居标签的概率为P_n，携带其他标签的P_t，其中$P_n>P_t$，携带新生标签的概率为P_w.

图4　知识激发问题网络模型

Fig.4　Knowledge-inspired problems network model

5)对于还未确定的标签空位，优先选择父标签的近邻标签，具体实现为随机选择步骤 4)中选定的标签为第二父标签，重复步骤 4)，优先选择第二父标签的直接邻居标签.问答平台上问题的核心讨论点可能不止一个，问题可能围绕多个不同的知识点展开，所以模型中的父标签也可能不止一个.

6)取与父标签距离为 3 及 3 以内的标签为父标签的近邻标签，所以步骤 5)可能重复实现，如果近邻标签的优先选择并未确定新生问题的所有携带标签，则由步骤 3)开始再选择父标签重复实现.此时视为新生问题由多个知识标签激发，而且问题围绕展开的知识点之间差异较大.

标签网络模型的生成算法所需要的各种参数将从实际网络的分析中得到.需要的参数包括新生问题携带标签数的概率分布、标签的增长速率、标签选择邻居标签的概率.

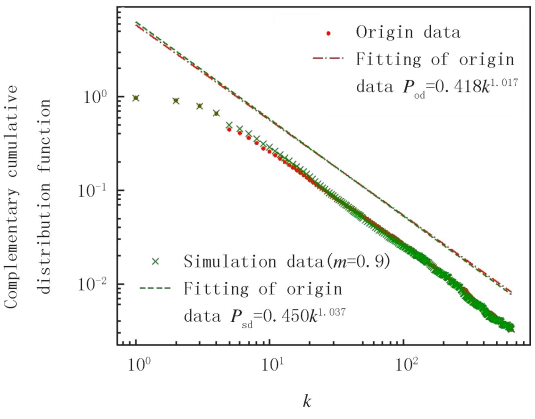
3.2　结果

根据每年子网络演化起点的真实数据构建仿真初始环境，按照知识激发问题模型调整仿真参数生成仿真数据，得到仿真网络并与实际标签知识网络对比分析.

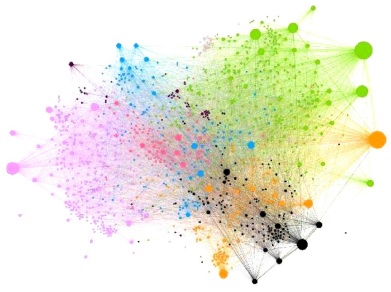
选择 2013 年仿真网络和实际网络的度值互补累计分布进行比对，结果如图 5(a)，仿真实现的网络度分布也符合幂律分布且两者之间差距在合理范围之内.对 2013 年仿真网络同样进行社团划分，得到的结果如图 5(b)，网络具有良好的社团结构.图 5 说明了模型得到的网络具有无标度特性且网络有良好的社团结构，说明构建的网络模型在节点度生长方面符合实际网络情况.

进一步分析标签激发问题能力的变化情况，也就是影响因子 m 的变化情况.仿真实现网络生长并比对每年份的实际子网络，发现网络的度值影响因子 m 逐年减小(见图 6).在网络发展初期，如 2011 年，分别设定度值影响因子 m 为 1.55 和 1.0 实现网络生长得到仿真 2011 年子网络并与实际网络对比度值累计分布.可以看出 m 取 1.55 时，仿真结果更贴合实际情况.再分别仿真其他年份子网络的生长，得到度值影响因子逐年减小.影响因子的减小对应了度值大的标签激发问题的能力的下降.知识标签激发问题的能力与其度值由超线性相关走向了亚线性相关.这样的网络生长特点揭示了知乎网站知识体系发展前期，热门的知识话题受到了更大比例的关注度，知识体系发展后期，热门话题的关注度得到了一定的平均化.

前面的实证分析可以得到标签新增连边存在优先选择直接邻居标签的情况.假设优先选择只存在于父标签的直接邻居标签而不考虑其近邻标签(与父标签距离 3 以内).仿真实现网络生长并与实际网络和原有模型仿真网络对比计算网络模块度，结果如图 7.当优先选择仅考虑父标签的邻居标签时，生成网络的模块度明显小于实际网络，而当优先选择存在于父标签的邻居标签和近邻标签时，生成网络的模块度更接近实际网络.说明网络的优先连接发生在距离较近的标签之间，即新生问题更倾向于选择携带父标签一定距离范围内的标签.



(a)度值互补累计分布图

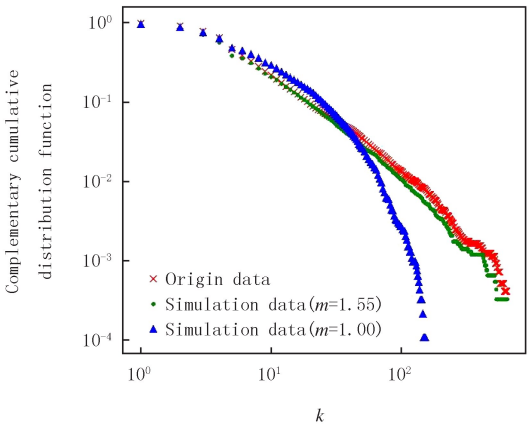


q=4.56

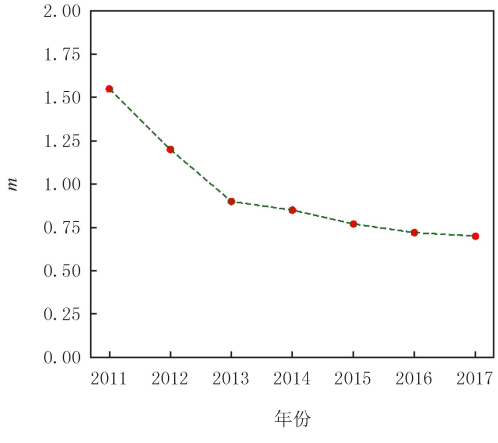
(b)网络社团结构

图5 仿真2013年子网络的度值互补累计分布和网络社团划分结果

Fig. 5 Complementary cumulative distribution of simulated network degrees and results of network community division in 2013



(a)度值互补累计分布图



(b)度值影响因子变化

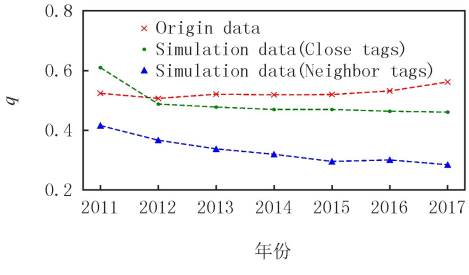
图6 度值影响因子变化

Fig. 6 Changes in the impact factor m

4 结论和讨论

本文以知乎网站知识标签网络为研究对象,收集了知乎网站 2011 年至 2017 年的问题以及标签数据,基于标签的共现关系构建了网络.从统计分析层面上得出知识标签网络具有复杂的网络结构,其度分布符合幂律分布,并且网络模块度较大,形成明显的抱团结构.

通过分析知乎知识标签网络的演化,得出网络新节点的出现速率会随网络节点的增加而放缓;节点新增连边选择连接已有连边的邻居节点的概率远高于随机,并且随着节点度的增大.这个概率也会增加.这些动态演化特点说明知乎知识体系中,新知识话题的产生受到已有知识话题的限制,有过关联的知识话题之间存在偏好连接.



叉点为实际网络,圆点为优先选择存在于近邻标签下仿真网络,三角为优先选择仅考虑直接邻居下仿真网络.

图7 不同优先选择范围下网络模块度变化情况

Fig. 7 Changes in network modularity in different situations

基于得到的网络动态演化特性,本文提出了知识激发问题的标签网络生长模型,模型能够很好地再现知

乎标签知识网络的网络结构.从问题出现的角度出发，提出了一个新的标签网络生长模型.模型以问题生成为网络演化周期，问题由父标签激发，标签激发问题的能力由其度值决定.新生问题在选择携带标签时优先选择父标签的近邻标签.根据模型得到仿真网络，对比实际网络，仿真网络能够再现知乎知识标签网络的无标度特性和社团结构，进一步分析得到知识标签激发问题的能力与其度值由超线性相关逐渐走向亚线性相关.相对以往的标签网络模型，本文的模型通过实证分析得到优先增长率在标签网络生长过程中的具体表现，因此建立的标签生长模型包含了更多演化信息.知识话题激发问题的模型从标签网络的微观增长出发，构建了具有复杂结构的标签网络，为研究其他知识网络的生成机制提供了一个新的视角.

参　考　文　献

[1] ULLAH A S.What is knowledge in Industry 4.0?[J].Engineering Reports,2020,2(8):e12217.

[2] 刘向，马费成，陈潇俊，等.知识网络的结构与演化：概念与理论进展[J].情报科学，2011，29(6)：801-809.
LIU X,MA F C,CHEN X J,et al.Structure and evolution of knowledge network:concept and research review[J].Information Science,2011,29(6):801-809.

[3] MEJIA C,KAJIKAWA Y.Emerging topics in energy storage based on a large-scale analysis of academic articles and patents[J].Applied Energy,2020,263:114625.

[4] CHANDRASEKHARAN S,ZAKA M,GALLO S,et al.Finding scientific communities in citation graphs:articles and authors[J].Quantitative Science Studies,2021,2(1):184-203.

[5] CHENG Q K,WANG J M,LU W,et al.Keyword-citation-keyword network:a new perspective of discipline knowledge structure analysis[J].Scientometrics,2020,124(3):1923-1943.

[6] LI H J,AN H Z,WANG Y,et al.Evolutionary features of academic articles co-keyword network and keywords co-occurrence network:based on two-mode affiliation network[J].Physica A:Statistical Mechanics and Its Applications,2016,450:657-669.

[7] FAGAN J,EDDENS K S,DOLLY J,et al.Assessing research collaboration through co-authorship network analysis[J].The Journal of Research Administration,2018,49(1):76.

[8] FU X,YU S D,BENSON A R.Modelling and analysis of tagging networks in Stack Exchange communities[J].Journal of Complex Networks,2019,8(5):cnz045.

[9] 裘江南，张美慧，宋晓玉.在线知识社区结构平衡对知识序化影响研究：以维基百科为例[J].情报学报，2017，36(3)：231-240.
QIU J N,ZHANG M H,SONG X Y.Research on the impact of OKC structural balance on the knowledge ordering:a case study of wikipedia[J].Journal of the China Society for Scientific and Technical Information,2017,36(3):231-240.

[10] 商宪丽，王学东，张煜轩.基于标签共现的学术博客知识资源聚合研究[J].情报科学，2016，34(5)：125-129.
SHANG X L,WANG X D,ZHANG Y X.Academic blog knowledge resource aggregations based on tag co-occurrences[J].Information Science,2016,34(5):125-129.

[11] 徐汉青，滕广青，安宁，等.基于模体的知识网络结构演化及其稳定性[J].图书馆学研究，2018(18)：82-90.
XU H Q,TENG G Q,AN N,et al.Structure evolution and stability of knowledge networks based on motifs[J].Research on Library Science,2018(18):82-90.

[12] 耿志杰，朱学芳，王文韬.情报学领域关键词同现网络结构研究[J].情报科学，2010，28(8)：1179-1182.
GENG Z J,ZHU X F,WANG W N.Research on the structure of key words co-occurrence network in information science[J].Information Science,2010,28(8):1179-1182.

[13] 滕广青，白淑春，韩尚轩，等.基于无标度与分形理论的层次知识网络原理解析[J].图书情报工作，2017，61(14)：132-140.
TENG G Q,BAI S C,HAN S X,et al.Analysis on the principle of knowledge network at level based on scale-free and fractal theory[J].Library and Information Service,2017,61(14):132-140.

[14] 潘旭伟，杨祎，王世雄，等.知识协同视角下 Wiki 知识网络的特性研究：以 Wikipedia 为例[J].情报学报，2013，32(8)：817-827.
PAN X W,YANG Y,WANG S X,et al.Empirical analysis of characteristics of Wiki knowledge network from the perspective of knowledge collaboration:a case study of Wikipedia[J].Journal of the China Society for Scientific and Technical Information,2013,32(8):817-827.

[15] 易明，曹高辉，毛进，等.基于 Tag 的知识主题网络构建与 Web 知识推送研究[J].中国图书馆学报，2011，37(4)：4-12.
YI M,CAO G H,MAO J,et al.Knowledge topic network construction and web knowledge push based on tag[J].Journal of Library Science in China,2011,37(4):4-12.

[16] SEGARAN T.Programming collective intelligence:building smart web 2.0 applications[M].[S.l.]:O'Reilly Media Inc,2007.

[17] 吴振宇，胡军，李德毅.社会标注系统幂律特性分析[J].复杂系统与复杂性科学，2014，11(2)：5-16.
WU Z Y,HU J,LI D Y.Analysis of the power law characteristics in social tagging systems[J].Complex Systems and Complexity Science,

2014,11(2):5-16.
[18] XIA H X,ZHAO X W,LIU H Y.Social tagging dynamics under system recommendation and resource multidimensionality[J].Journal of Systems Science and Systems Engineering,2016,25(3):271-286.
[19] 冯鑫,胡姝慧,李佳培,等.基于复杂模体的标签网络演化特征研究:以问答学习社区知乎为例[J].情报科学,2020,38(9):56-62.
FENG X,HU S H,LI J P,et al.Research on evolution characteristics of tagging networks based on complex motif:the case of Q & A learning community Zhihu[J].Information Science,2020,38(9):56-62.
[20] BARABÁSI A L,ALBERT R.Emergence of scaling in random networks[J].Science,1999,286(5439):509-512.
[21] KRAPIVSKY P L,REDNER S,LEYVRAZ F.Connectivity of growing random networks[J].Physical Review Letters,2000,85(21):4629-4632.
[22] GOLOSOVSKY M,SOLOMON S.Stochastic dynamical model of a growing citation network based on a self-exciting point process[J].Physical Review Letters,2012,109(9):098701.
[23] JEONG H,NÉDA Z,BARABÁSI A L.Measuring preferential attachment in evolving networks[J].Europhysics Letters(EPL),2003,61(4):567-572.
[24] 韩仪,冯鑫,周金连,等.知识标签网络生成机制研究[J].电子科技大学学报,2021(2):294-302.
HAN Y,FENG X,ZHOU J L,et al.The generation mechanism of label network[J].Journal of University of Electronic Science and Technology of China,2021(2):294-302.
[25] BLONDEL V D,GUILLAUME J L,LAMBIOTTE R,et al.Fast unfolding of communities in large networks[J].Journal of Statistical Mechanics(Theory and Experiment),2008(10):10008.

Study on the formation mechanism of Zhihu tagging network

Huang Tao[1a], Wang Shengfeng[1b], Wu Ye[2], Zhang Peng[1a], Xiao Jinghua[1a]

(1a. School of Science; b. School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China; 2. Computational Communication Research Center, Beijing Normal University, Zhuhai 519087, China)

Abstract: Knowledge networks are important for exploring knowledge development, thus it is important to study the statistical characteristics and formation mechanism of knowledge networks. A kind of knowledge network, tagging networks has received attention from researchers in recent years. However, there is still a lack of research on the growth mechanism of tagging networks. In this paper, we construct tagging networks based on the tagging data of Zhihu Q & A platform, and statistically analyze its static statistical characteristics and evolutionary characteristics. In order to understand the complex structure of the tagging network, we propose a formation model for the network, which assumes that new questions are inspired by knowledge tags and the ability of knowledge tags to inspire questions is positively related to their degree. The simulation results show that the model can well reproduce the statistical properties and the association structure of the tagging network. This paper reveals the dynamic evolutionary characteristics exhibited in the growth process of the tagging network. The formation model of the tagging network based on the empirical results is enlightening for understanding the formation of other knowledge networks.

Keywords: tagging networks; question and answer community; evolutionary model; power law distribution

[责任编校 杨浦 刘洋]